Assignment #4

Jingye Wang

October 4 2016

13.2

Models for adjusting individual ratings: a committee of 10 persons is evaluating 100 job applications. Each person on the committee reads 30 applications (structured so that each application is read by three people) and gives each a numerical rating between 1 and 10.

Part A

It would be natural to rate the applications based on their combined scores; however, there is a worry that different raters use different standards, and we would like to correct for this. Set up a model for the ratings (with parameters for the applicants and the raters).

$$Yi = \alpha j[i] + \beta xj + \epsilon i$$

$$\alpha j[i] = \gamma_0^\alpha + \gamma_1^\alpha rj + \eta_j^\alpha, for\ j = 1, \ldots, 120$$

Part B

It is possible that some persons on the committee show more variation than others in their ratings. Expand your model to allow for this.

$$Yi = \alpha j[i] + \beta j[i] xi + \epsilon i$$

$$\beta j \sim N\left(1, \sigma_\gamma^2\right),\ \ for\ j = 1, \ldots, 120$$

$$\alpha j[i] = \gamma_0^\alpha + \gamma_1^\alpha rj + \eta_{j,}^\alpha, for\ j = 1, \ldots, 120$$

13.4

Models with unequal variances: the folder age.guessing contains a dataset from Gelman and Nolan (2002) from a classroom demonstration in which 10 groups of students guess the ages of 10 different persons based on photographs. The dataset also includes the true ages of the people in the photographs. Set up a non-nested model to these data, including a coefficient for each of the persons in the photos (indicating their apparent age), a coefficient for each of the 10 groups (indicating potential systematic patterns of groups guessing high or low), and a separate error variance for each group (so that some groups are more consistent than others).

import data

```
age_data_raw <- read.csv('age_guessing.csv', header = T, row.names = 'Group', sep = ';')
```

convert dataset

```
age_data <- as.data.frame(t(age_data_raw[,c(1:10)]))
age_data$photoid<-seq.int(nrow(age_data))
Trueage_data <- age_data[,c(11,12)]
age_data <- age_data[,-11]
age_reshape <-reshape(age_data, direction="long", varying=list(names(age_data)[1:10]),
v.names="error", idvar=c("photoid"), timevar="groupid")
age_reshape <- cbind(age_reshape,Trueage_data)
age_reshape <- age_reshape[,-5]
colnames(age_reshape) <- c('photoid', 'groupid','error', 'trueage')
age_reshape$guessedage <-age_reshape$error + age_reshape$trueage
```

build the model

```
model_0 <- lmer(formula = guessedage ~ 1 + (1 | groupid) + (1 | photoid), data = age_res
hape)
summary(model_0)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: guessedage ~ 1 + (1 | groupid) + (1 | photoid)
##    Data: age_reshape
## 
## REML criterion at convergence: 590.6
## 
## Scaled residuals: 
##     Min      1Q  Median      3Q     Max 
## -2.1875 -0.5950 -0.1133  0.5814  3.6267 
## 
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  groupid  (Intercept)   0.1422  0.3771 
##  photoid  (Intercept) 368.9156 19.2072 
##  Residual              12.8933  3.5907 
## Number of obs: 100, groups:  groupid, 10; photoid, 10
## 
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)   39.940      6.086   6.563
```

Based on the result, the std.dev of intercept for groupid is 0.3771, the std.dev of intercept for photoid is 19.2072. Compared to the std.dev of residual (3.5907), we can conclude that photoid are more suitable to be used in building a multi-level model.

13.5

Return to the CD4 data introduced from Exercise 11.4.

import data

```
hiv_data <- read.csv ("allvar.csv", header= T)
hiv_data$time <- hiv_data$visage - hiv_data$baseage
hiv_data$time_round <- as.factor(round(hiv_data$time, digits = 1))
```

Part A

Extend the model in Exercise 12.2 to allow for varying slopes for the time predictor.

```
model_1 <- lmer(sqrt(CD4PCT) ~ time + factor(treatmnt) + baseage + (1 + time | newpid),
data= hiv_data)
summary(model_1)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: sqrt(CD4PCT) ~ time + factor(treatmnt) + baseage + (1 + time |
##     newpid)
##    Data: hiv_data
##
## REML criterion at convergence: 3107
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -5.0998 -0.4057  0.0174  0.4030  5.0157
##
## Random effects:
##  Groups   Name        Variance Std.Dev. Corr
##  newpid   (Intercept) 1.8464   1.3588
##           time        0.3374   0.5808   -0.04
##  Residual             0.5145   0.7173
## Number of obs: 1072, groups:  newpid, 250
##
## Fixed effects:
##                   Estimate Std. Error t value
## (Intercept)        5.10850    0.18594  27.474
## time              -0.35258    0.06763  -5.214
## factor(treatmnt)2  0.15952    0.18137   0.880
## baseage           -0.12423    0.03971  -3.128
##
## Correlation of Fixed Effects:
##             (Intr) time   fct()2
## time        -0.114
## fctr(trtm)2 -0.463  0.010
## baseage     -0.729 -0.013 -0.004
```

Based on the result, the std.dev of slope for time is 0.5808, the std.dev of intercept is 1.3588. Compared to the std.dev of residual (0.7173), we can conclude that time is not a suitable variable to be used in building a multi-level model based on patient id.

Part B

Next fit a model that does not allow for varying slopes but does allow for different coefficients for each time point (rather than fitting the linear trend).

```
model_2 <- lmer(sqrt(CD4PCT) ~ time + factor(treatmnt) + baseage + ( 1 | time_round), da
ta = hiv_data)
summary(model_2)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: 
## sqrt(CD4PCT) ~ time + factor(treatmnt) + baseage + (1 | time_round)
##    Data: hiv_data
## 
## REML criterion at convergence: 3976.8
## 
## Scaled residuals: 
##     Min      1Q  Median      3Q     Max 
## -3.1993 -0.5336  0.1416  0.6967  2.9442 
## 
## Random effects:
##  Groups     Name        Variance Std.Dev.
##  time_round (Intercept) 0.000    0.000   
##  Residual               2.365    1.538   
## Number of obs: 1072, groups:  time_round, 19
## 
## Fixed effects:
##                   Estimate Std. Error t value
## (Intercept)        4.96386    0.11334   43.79
## time              -0.18327    0.09860   -1.86
## factor(treatmnt)2  0.31807    0.09434    3.37
## baseage           -0.10371    0.02081   -4.98
## 
## Correlation of Fixed Effects:
##             (Intr) time   fct()2
## time        -0.495              
## fctr(trtm)2 -0.402  0.015       
## baseage     -0.646 -0.033  0.023
```

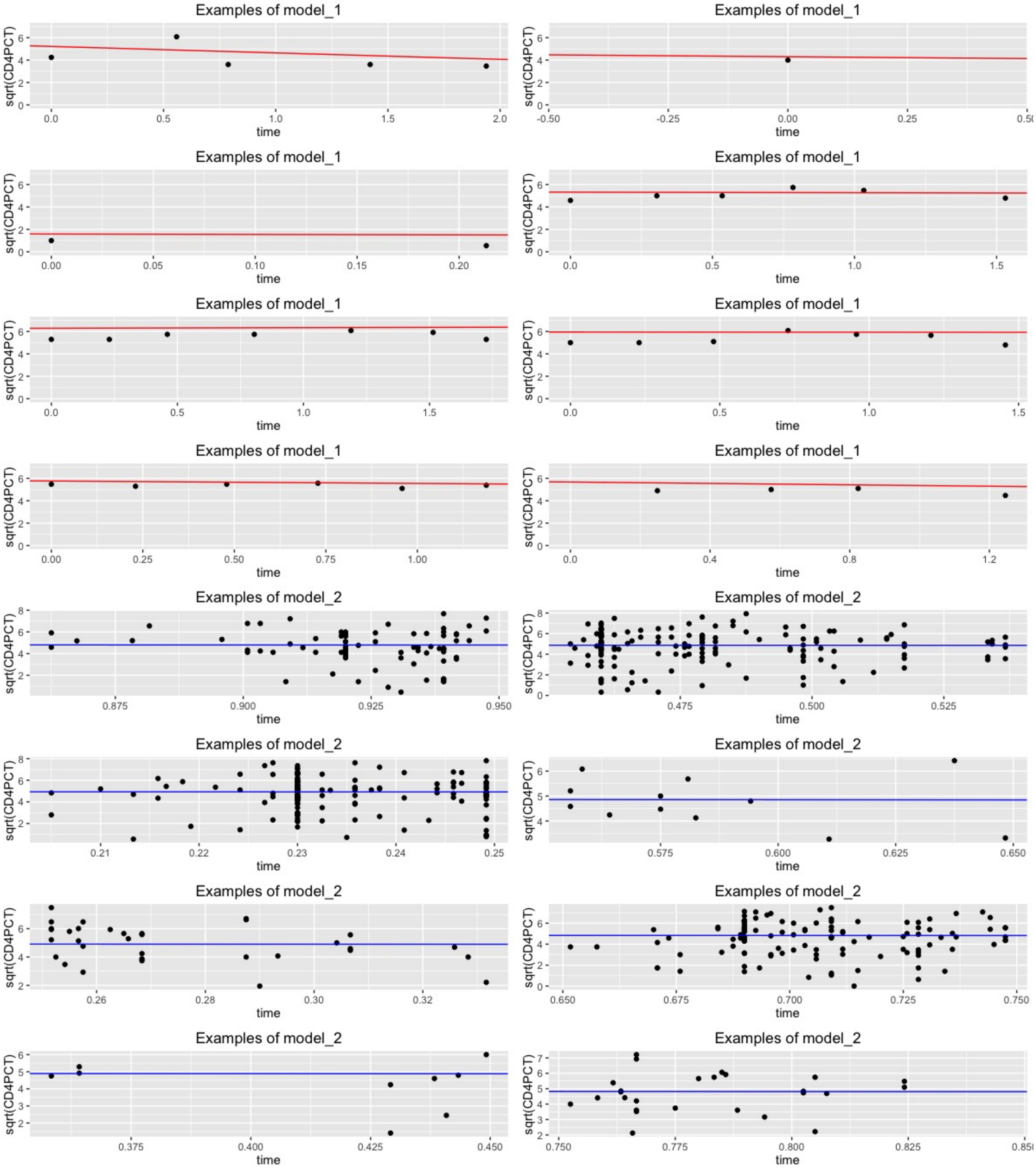
Based on the result, the std.dev of intercept is 0. Compared to the std.dev of residual (1.538), we can conclude that time_round is not a suitable to be used in building a multi-level model.

Part C

Compare the results of these models both numerically and graphically.

```
coef_df_1 <- data.frame(coef(model_1)$newpid)
colnames(coef_df_1) <- c('Intercept', 'time', 'treatment', 'baseage')
coef_df_2 <- data.frame(coef(model_2)$time_round)
colnames(coef_df_2) <- c('Intercept', 'time', 'treatment', 'baseage')
plot_list = list()
for (i in 1:8){
p <- ggplot(data = hiv_data[hiv_data$newpid==i, ], aes(time, sqrt(CD4PCT))) + geom_point()
 + geom_abline(slope = coef_df_1[i, 2], intercept = coef_df_1[i, 1], colour = 'red') + g
gtitle('Examples of model_1 ') + ylim(0,7)
plot_list[[i]] = p
}
multiplot(plot_list[[1]],plot_list[[2]],plot_list[[3]],plot_list[[4]],plot_list[[5]],plo
t_list[[6]],plot_list[[7]],plot_list[[8]],cols=2)

plot_list = list()
for (i in 2:9){
i = i/10
p <- ggplot(data = hiv_data[hiv_data$time_round==i, ], aes(time, sqrt(CD4PCT))) + geom_p
oint() +  geom_abline(slope = coef_df_2[i*10, 2], intercept = coef_df_2[i*10, 1], colour
 = 'blue') +  ggtitle('Examples of model_2')
i=i*10
plot_list[[i]] = p
}
multiplot(plot_list[[9]],plot_list[[2]],plot_list[[3]],plot_list[[4]],plot_list[[5]],plo
t_list[[6]],plot_list[[7]],plot_list[[8]],cols=2)
```



Compare Redisual: Model_1 has a residual of 0.7173 and Model_2 has residual of 1.538. The std.dev of intercept in model_1 is 1.3588 and the std.dev of intercept in model_2 is 0. These suggest that model_1 explained sqrt(CD4PCT) better.